Data identity and perspectivism

Synthese
https://doi.org/10.1007/s11229-020-02824-8

Data identity and perspectivism

Franklin Jacoby[1]

Received: 4 May 2019 / Accepted: 4 August 2020

Abstract
This paper uses several case studies to suggest that (1) two prominent definitions of data do not on their own capture how scientists use data and (2) a novel perspectival account of data is needed. It then outlines some key features of what this account could look like. Those prominent views, the relational and representational, do not fully capture what data are and how they function in science. The representational view is insensitive to the scientific context in which data are used. The relational account does not fully account for the empirical nature of data and how it is possible for data to be evidentially useful. The perspectival account surmounts these problems by accommodating a representational element to data. At the same time, data depend upon the epistemic context because they are the product of situated and informed judgements.

Keywords Perspectivism · Data · Evidence · Representation · Relational account

1 Introduction

What are scientific data? There are two main answers. One influential answer, first defended by Bogen and Woodward (1988), is that data are representational. They represent in virtue of being records produced by reliable experiments. Data provide empirical evidence and, as such, are free from theoretical assumptions and determined, in crucial ways, by nature. They are also stable, meaning their identity does not change even if theoretical or experimental practices change.

Another answer, more recently defended, is the relational account (Leonelli 2016). Data are defined principally by their use as evidence. Consequently data identity depends upon the particular inquiry in which they feature and their identity changes as the inquiry changes. Different inquiries put different demands upon data and, to meet

✉ Franklin Jacoby
frjacoby@fastmail.com

[1] Cherryfield, USA

Published online: 17 September 2020

these demands, data change identity. Because these demands are many and varied, data identity changes often.

Given these two seeming incompatible options, how should we define scientific data? Do data have a changing identity and, if so, what precipitates identity change? In trying to clarify the role of data in science, this paper will strike a middle option between the representational and relational accounts. I will call this third view a perspectival account, which is commited to two claims about data: (1) data identity changes much less frequently and easily than the relational account suggests because data are representational; and (2) data identity is not completely stable because data depend upon distinctions that scientists make. This dependence can be helpfully understood by appeal to perspectivism.

In Sect. 2 I discuss why the relational account, as I have presented it, provides a foil to explicating a view with stronger representational commitments. Section 3 develops a case study that suggests data have some representational element. Section 4 shows what a representational view that is sensitive to historical cases might look like.

2 Representational and relational accounts of data

The representational view of data (Latour 1999; Rheinberger 2011; Bogen and Woodward 1988; Bogen and Woodword 2003) is the view that knowledge claims are grounded on a largely theory-free and empirical contribution from data. Data are theory-free in the sense that they are not the kind of objects that scientists attempt to explain using theory, nor can they be derived from or predicted by theory (Woodward 1989, p. 394). As well as being theory-free, data are mostly independent from the epistemic context, i.e. independent of the actions, beliefs, and other epistemic features of the scientists who collect or use them. I say "mostly" because the representational view does make room for some of these considerations, but they allocate them to "noise," that is, factors that obscure the causal origin of the data (Woodward 2010, p. 793). Data, when reliably produced, provide a signal and the causal origin of that signal is a phenomenon. That is why, Woodward thinks, data can be used as evidence: because they are causally produced by phenomena. These henomena are not, however, the only causes acting on data (*Ibid*). Any causal factors that are not the phenomenon of interest count as noise. For example, if I am interesting in the melting point of lead and I measure a series of melting lead samples, I should get consistent results. The consistency, the signal, is due to the phenomenon (melting lead), but I am likely to have some variation in my data. The variation (noise) is likely to due to extraneous factors, such as impurities in the sample of lead, incorrect positioning of the thermometer, a malfunctioning thermometer, failure to measure temperature at the correct moment, and so on. One crucial feature of this view is that the causal role the scientist plays is restricted to the noise. Data are largely independent of the epistemic context in that the causal origin of any data set is free from human interference.

Scientists record data and, because phenomena are the causal origin of data, the data can serve as unchanging evidence against which theories and models can be compared, or upon which theories and models are built. Data serve as the empirical arbitrator by being evidence that supports more theoretical claims about phenomena, phenomena

such as the melting point of lead, neutrinos, black holes, aggressive behaviour, or evolutionary traits (Bogen and Woodward 1988).

Throughout their collection and use, data are stable because they are causally connected, sometimes through long causal chains, to stable phenomena. We can see Woodward making this point in two stages. First, phenomena are stable:

> Phenomena, as I shall use the term, are relatively stable and general features of the world which are potential objects of explanation and prediction by general theory. (Woodward 1989, p. 393).

Because phenomena are stable features of the world, it stands to reason that anything causally produced by those stable features is at least partly stable, which in this case is data.[1] Woodward makes this point a little later in his (1989, p. 404):

> It is very common to understand in principle how a phenomenon plays a causal role in the production of a certain body of data, without being in a position to extract reliable information from that data regarding the phenomenon in question.

We can see from this quotation that data and phenomena are causally linked and that this link exists regardless of whether scientists are "in a position" to know anything about that link. This suggests to me that the representational view treats data, as well as the phenomenon that caused them, as stable, i.e. possessing an unchanging identity in the face of epistemic change.

Leonelli (2009, 2016, 2015) rejects the representational view and defends a relational account. This second account claims data are material artefacts whose identity is determined by their evidential use. After describing how this view is motivated, I will argue that accounts that do not treat data as at least partly representational—face two issues: the problem of identity and the problem of data stability.

Leonelli argues that data must be able to travel and that they must be evidential; these two considerations require substantive identity changes in data (Leonelli 2012, 2009, 2013, Leonelli 2016). Here is an excerpt from Leonelli's work where she discusses this change in identity, couched in terms of stability:

> What I do not share [with the representational view of data] is the emphasis on stability. When travelling from their original context of production to a database, and from there to a new context of inquiry, biological data are anything but stable objects. (Leonelli 2016, p. 5).

Data must be transported to be used: they are not used right at the time and location of collection. To make data suitable for movement and use, they must be formatted, classified, organized with meta-data, and filed for later use. Scientists make these material changes so the data can be put to new and different evidential uses. Leonelli also writes:

[1] It is worth nothing that the representational view does not take data to be as stable as Phenomena. Woodward (2010, p. 793) notes that data are affected by local idiosyncrasies of experiment, for example. However, the causal origin remains the phenomena and any data reliably produced from that phenomena will share a shared relative stability. Also, once collected, data do not change.

> Within this framework, it is meaningless to ask what objects count as data in the abstract, because data are defined in terms of their function within specific processes of inquiry (*Ibid*. p. 7).

This passage suggests that data are defined by their use, or the "role they are made to play" (Leonelli 2016, p. 78). Leonelli's account is relational because it rejects the consideration of a datum independently of the context in which it is used. Consequently, what are data to one scientist in one context of inquiry may be different data to a scientist in a different context, or not data at all. This is the particular feature of relational accounts, namely that data are "defined in terms of their function within specific processes of inquiry" that I want to critically engage with in the rest of this paper and somehow mitigate by arguing that the representational view might in fact have some important insights, even though we do not want to reject the importance of examining data within the epistemic contexts in which they are used. Let us then consider in more detail what data are according to Leonelli's account:

> […] any object can be considered as a datum as long as (1) it is treated as potential evidence for one or more claims about phenomena, and (2) it is possible to circulate it among individuals" (2015, p. 2).

A set of records are data when those objects function as evidence and when the set can be transported. Data are defined by their functioning as potential evidence *and* by virtue of their physical form. Because data are defined by these two criteria, a set of data does not merely acquire or lose properties or characteristics when its use or form changes, the set of data actually ceases to be data or becomes different data. *The identity of the data set has changed.* Data, this view seems to suggest, are unstable because what data are depend on purposes of specific agents that want to make evidential claims about some phenomena. As soon as different epistemic communities have different purposes in mind, what was once a data set might no longer count as such.

My purpose in what follows is to address the problem of data instability by suggesting that relational accounts leave room for a further discussion concerning what data are in addition to their evidential role. There are two issues to address concerning the relational account: (I) is use-as-evidence sufficient in identifying data (discussed in 2.1.)? and (II) is the materiality of the data important in establishing identity (discussed in 2.2.)?

2.1 Potential evidence

I here suggest (I) is problematic for two reasons. First, data are not the only source of evidence in science. A model or simulation, for example, might provide evidence that a hurricane will strike a particular place at a particular time, but both these forms of scientific evidence are distinct from scientific data.

Leonelli acknowledges that the products of models and simulations can indeed be evidential, but for her they are therefore data (2015, p. 817). But this is strange because in the case of hurricane predictions, there are two very different pieces of information involved. Such models and simulations take something that is known or collected

(inputs) and create predictions (outputs). To call both data—which I believe relational accounts must because they are both evidential—misses an important epistemic difference between data that can be used as a basis for modelling and the output, which is something in which we have much less epistemic confidence. In essence, there is a distinction between data—things we collect or record—and predictions. Defining data in terms of evidence fails to respect this distinction.

But what is evidence anyway? There is, I think, general agreement between representational and relational views. Both Bogen and Woodward (1988, p. 304) and Leonelli (2013, p. 505) argue data are evidence because they provide support for claims, typically claims about phenomena. Both (*Ibid.*) also suggest an important element of how data provide evidence emerges from patterns in data sets. Depending on how and if one wanted to characterize patterns and support, the notion of evidence is quite permissive, which suggests that it is reasonable to think a number of different forms of evidence (of which data is one) are possible.

This further suggests, together with an appreciation of models and simulations, data are not the *only* evidence. How then should they be distinguished from other forms of scientific evidence (say simulated evidence)? Or, differently put, can simulated evidence—the outputs of a simulation—count as evidence under the relational account? And if so, is not there a risk that our theorizing here is too course to capture what count as data and therefore obscuring important epistemic distinctions?

Second, it is unclear how data can be used as potential evidence. To use data as evidence seems straightforward, but "potential evidence" is presumably not a use because any such use would just be a straightforward case of using data as evidence. This suggests potential use is more like an attitude scientists have toward a set. If it is this attitude that defines data, then we have a problem.

The problem stems from how to identify a set of data based on an expectation. The expectation that a set of data will be evidentially useful seems almost intuitive; the products of experiment will be useful in this way, otherwise why conduct experiments? It also means that scientists can perform experiments and collect data without necessarily needing to know what claims that potential evidence will support. Finally, identifying data as potential evidence means that what scientists collect in an experiment—or another research setting—can be useful evidentially to other scientists in other places who are investigating different research questions.

The problem now is this: what is this potential evidence produced through experiment or observation? Is it data—and hence potential evidence—or is it just a record? Leonelli seems to suggest that potential evidence, i.e. an expectation about the usefulness of data, makes objects data. But if that is the case, then how a set of data is used evidentially for any particular claim is not an identity condition. In other words, one's records are data regardless of the particular claim that those data support. This in turn suggests that how data are used as evidence does not change the identity of those data. It also suggests a set of objects that are potential evidence—and therefore a set of data—do not change identity when they are transported or put to different uses, provided there remains an expectation about evidential value. If the use of data does not determine their identity, then data are not as relational as it may seem. This is a problem for some of the claims Leonelli makes, specifically when she writes that

> […] the same objects may or may not be functioning as data, depending on which role they are made to play (p. 817). (Leonelli 2015, p. 817)

This suggests that objects, or records, are or are not data depending on their evidential use, or the "role they are made to play." But if data are potential evidence, then surely the particular role cannot be so significant.[2]

These two issues—defining data so as to include too many types of evidence and appealing to potential evidence—show that data should not be defined solely in terms of evidence: something further must be said about what data are if we are to understand how they can be used. My point here is not that data are *not* evidence, but that in order to address why data can function as evidence, there is a logically prior question about what they are.

2.2 Data materiality

But perhaps this is too hasty. Leonelli may have anticipated some of these objections and may have more of a story to tell about what defines data. For she compares data to biological individuals, which also have some kind of identity through continuous change over time, like a succession of states (2016, p. 82). This comparison preserves the intuition that something about data does persist across time and space, even though what persists continuously changes. If this is right, then data can be connected to their collection and to previous uses, giving scientists motivation for treating them as potential evidence, while at the same time preserving a relational account.

This analogy only works loosely and it only works if we consider the materiality of data important for their identity. I do not have space to properly address this response, but I will briefly discuss why it should not satisfy us. First, biological individuals may be too unlike data. We might think, for instance, that reproduction, evolutionary relationships, birth, and death are important determinants of biological individuality. These considerations are not, however, relevant for data. Second, biological individuals, to the extent they change, do so materially; this is what the analogy hangs on. I am sceptical that materiality is important for data identity.

One reason for thinking the material change in data can be overplayed stems from the high level of stability data often have. This is *the problem of instability*: data must have some stability to function as evidence. This is exemplified in cases where changes in data form are symmetric: for example, cases where we can take digital information and write it down on a data sheet, then enter that information on a computer, thus getting the same material object that we started with. Transformation that allows for a return to the original form in this way does not seem very substantive since data can often be moved and transformed without *loss of information*. If the information were lost, then change back to the original would not be possible.

The requirement that data be stable is also exemplified when data are moved and transformed, especially when scientists critique or respond to one another. Consider as an example the historical and philosophical work that Allan Franklin (1981) conducted on the electron's charge, building on Millikan's oil drop experiment. In the early

[2] I am grateful to an anonymous reviewer for clarifying my thinking on Leonelli's account and help in developing this argument.

20th century, Robert Millikan was interested in precisely measuring the charge on an electron and famously measured the falling rate of electrically charged oil drops to do so. Based on how fast the drops fell, he was able to calculate the size of an electron charge. He did not, however, publish all his data, only a selection. Franklin (*Ibid.*) revisited Millikan's notebooks to see, among other things, how Millikan's conclusion might have differed if all data were included in the calculations. This is where my point about transformation arises. Millikan recorded his data in a notebook using a pencil. His notebooks were later photographed and the photographs were stored as microfilm in the Millikan Collection at the California Institute of Technology. Franklin obtained digital copies of these microfilms (or at least some versions of his published paper used digitized versions of the microfilms). It seems to me that the data underwent extensive transformations before Franklin could verify Millikan's results. They began as pencil parks in a notebook and ended as bytes on a hard drive several decades later. Despite these extensive transformations, it would be odd to say that Franklin was *not* working with Millikan's data; the entire purpose of Franklin's work was to re-examine the data to determine whether some data points that Millikan omitted affected the results. This suggests that, despite some great material and contextual changes, Millikan's data did not change.

Leonelli's account of data has an insight here by noting the significance of the form data must take to use them in certain ways. However, it would seem that the identity of the data (in terms of their informational content) persists through this material change. Specifically, the persisting information, in the Millikan case, is the record of the following:

> The notebooks contain observations on 175 drops along with voltage and chronoscope corrections and measurements of the density of clock oil. (Franklin 1981, p. 187)

The notebooks contained information about the charge (voltage) on each drop as well as time corrections and the density of the oil. They are the records of the measurements and observations that Millikan made and consisted, in this case, of a table of numbers with labelled columns. This information seems to me the same regardless of whether it is in a notebook, Franklin's hard drive, or transcribed from my own computer to my notebook. I take it a relational account is committed to the idea that the data on the hard drive are different from the data in the notebook and this is puzzling.

These material changes may seem too trivial to worry the relational account, but they are actually quite substantial. Once microfilmed, Millikan's notebooks were archived with other material from his life and curated. There is even a published guide to assist the researcher in navigating the microfilm archive (Goodstein et al. 1977). Extracting data from these notebooks was no easy task.

A relational view of data might want to give due consideration to elements of data practices, such as those illustrated by this electron charge example. Indeed, any account should do this and the view I defend below seeks to accommodate this consideration. But an account must also be able to explain how different scientists across time with different interests and using different tools could nonetheless study the same phenomenon and I do not believe relational accounts have an explanation for this as yet.

Franklin and Millikan were both interested in the same oil drops and the same charges on those oil drops. Their research interests were slightly different: Millikan wanted to calculate the charge of an electron and Franklin wanted to determine whether Millikan made no important omissions. Both of these different research interests required the same data set. It is difficult to make sense of how they could have their respective research interests and pursue them, unless they were working with the same data. This case suggests that the data, the records that began in a notebook, provided a link between two scientifically-minded researchers and a set of oil drops. This conclusion gets us out of the problem of stability: data are stable enough to support a variety of research interests and this stability stems from the fact that they are records, not just evidence.

Relational accounts do have resources to discuss the relationship between data and the world. One possible strategy may be to use meta-data. Leonelli discusses (2016, pp. 189–90) the importance of curatorial work in packaging data. Such work involves recording what kind of experiment and recording techniques were used in producing the data. Such meta-data are important for communicating how data were collected, what instruments were used, who made the record, and under what conditions, etc. So perhaps meta-data can explain why scientists should expect data to serve as evidence.

Without denying the importance of meta-data, one might still worry about how meta-data make a set of data *that set* and different from another set. If I read Leonelli correctly, meta-data are primarily important for evidential, but not identity, reasons. By allotting data-collection and experiment to meta-data, this relational account suggests data identity is not *primarily* affected by how data are produced. Assigning meta-data this secondary role is reasonable; after all, meta-data must be recorded in addition to recording data, but I contend that a set of data would have a life of its own even if one neglected to record the meta-data (though they may be evidentially not useful).

To sum up, I have argued that the ontology and use of a data set leave open questions about data identity. I noted that data identity, if data are potential evidence, is not affected affected by their use. Nor is the materiality of data important for determining their identity. This leaves room, I believe, for a treatment of what data are *qua* data, independent of evidence.

3 A new star

The history of science provides a rich source for thinking about what data are and how they change through time. Astronomy, because of its particularly long history, is an especially rich resource. If an account of data is to do justice to the great change and continuity in a scientific tradition, then astronomy is the paragon for judging that account and from which an account might be built. In this section, I discuss a case study from astronomy that suggests we think about data in a different, perspectival way. I discuss what this perspectival view looks like in Sect. 4.

At the end of the 2nd century A.D., astronomers in China recorded a "guest star." They recorded when the "star" appeared, when it disappeared, and its rough location in the night sky. These observations made their way into into a number of historical texts (see Wang et al. 1997). This record has since sparked several contemporary studies

that attempt to make sense of what this "new star" might be in contemporary terms. First, Clark and Stephenson (1977)) made an attempt to use this early observation. Thorsett (1992) and Green and Stephenson (2008) published further discussions of the issue. There are two challenges that all of these studies faced: one easy, one very challenging. The first is that these ancient observers did not write English and the second is that "new star" might mean several things. Making use of that ancient observation requires overcoming those two issues with two corresponding tasks. The first is a straightforward translation of ancient Chinese into English; this was the easy task. If you know the relevant languages, translation is straightforward. But translation alone does not establish what it was the ancient astronomers saw in contemporary terms and this second task was the harder. To do this, Thorsett (1992) suggested that the "new star" might be the supernova MSH15-52, which was in contrast to a set of supernovae suggestions from Clark and Stephenson (1977). Thorsett went about arbitrating between these possibilities using a number of methods. For one thing, a contemporary pulsar may have originated from supernova MSH15-52. Pulsars are small stars that can be detected using their radio emissions. Sometimes they are produced by supernovae. Thorsett estimated how old that pulsar is (which gives an indication of when the supernova occurred). This timing estimate seemed to match the Chinese observation. Thorsett also investigated three further considerations: where supernova MSH15-52 was likely to be in the night sky; how bright it was; and whether these estimates matched historical observation, or at least did not obviously conflict with what remains of the historical record. The result of these estimates and comparisons was that Clark and Stephenson's suggested supernovae did not fit well with the record, but MSH15-15 did.

The whole purpose of these studies was to put more specific constraints on contemporary theories. If modern astronomers could determine precisely when the supernova occurred, contemporary astronomers could use that information to test predictions more precisely.

This example illustrates a change in data. The modern astronomers did not simply take the historical record and "interpret" the observation in contemporary terms. If it was a matter of interpretation, then the record itself should provide sufficient information for a translation. An interpreter (translating French into English) need only hear the French phrase to translate it into English: no other research or information is required. This astronomy example is not so simple. The ancient Chinese did not discriminate between comets, supernovae, and some other astronomical phenomena. From their perspective, bright objects in the sky were all "stars," which was a completely reasonable judgement given the epistemic context, i.e. given their understanding of what we call astronomy. However, contemporary scientists make much finer distinctions between bright objects in the sky and "star" is too vague a term to constrain contemporary theories. So rather than just interpret the record in a new way, modern astronomers had to use the historical description of the event in conjunction with contemporary data and knowledge of astronomy to determine what event the Chinese astronomers observed. This process is very much like identification, which requires an understanding of how to determine the identity of an object, which is precisely what Thorsett knew and what he did.

This data change is not merely a change in interpretation and for two reasons. First, the data in this example presuppose content. Thorsett and other contemporary astronomers were not just using raw numbers both in working with ancient texts and in working with the data they extracted: they worked with records *of* something that was observed. Initially those records were of a "new star" and after contemporary research and re-identification, they were records of a supernova. In order to place constraints on contemporary theory, Thorsett's interpretation of the data first *required* that the data be records of very specific things, such as a supernova. Otherwise the data would have been useless. This suggests an essential feature of data is that they be records *of* something. I'll return to this point shortly.

Second, this example illustrates how once re-identified, the new data cannot be used in a straightforward way by the ancients. Ancient Chinese astronomers did not recognize supernovae and would be unable to use an observation about a supernova without first learning about modern astronomy and the taxonomic distinctions it recognizes, which would be highly anachronistic. This is distinct from interpretation, which is a symmetric relation (a phrase in French can be translated into English and then back into French).

Despite the profound changes going on in this case, it is important to note that there is a sense in which there is something that persists between the ancient Chinese records and the contemporary data. Otherwise, what would be the connection between the contemporary research and that event thousands of years ago? Surely there is one. The thing that persists is some courser grained description of the event, such as "small bright light that appears at such-and-such time and location." Although "small bright light" is not equivalent to "new star" nor to "supernova," even setting translating texts aside, such a course description could guide contemporary astronomers to historical records.

The course description also illustrates one way we could track the origins of Thorsett's data in this particular example to those who made the observation and why a historical event is of interest to the contemporary scientist. The course description is not, however, sufficient on its own for the contemporary astronomer. This is because the researcher needs one object to satisfy the course description—a specific supernovae—and there are too many objects or events that could satisfy the course description. The historical criteria and the criteria supplied by the course description are not appropriate for the contemporary context. The course description allows us to say something about what two very different sets of data share, but it does not, on it's own, capture what the historical scientists were doing, nor does it capture what contemporary scientists must know in order to pursue their research.

This description of re-identification may resemble Leonelli's "re-contextualization" (Leonelli 2016, p. 189). Despite the similarity, they differ in important ways. Leonelli notes that data are often packed and described in such a way as to allow them to be transported and later used (called de-contextualization). When scientists retrieve de-contextualised data, they then must "re-contextualise" them by situating the data, which involves classifying, describing, and forming them materially so as to provide evidential value for the particular inquiry.

This is substantially different from my view in two ways. First I believe this case study shows that data, to be used as evidence, must first be records of something,

regardless of how an individual scientist has used them, provided the *perspective is the same*. If the perspective is the same, then the observations that the data are records of are also the same, i.e. they are unchanged. Section 4 provides further clarification on what a perspective is.

It is necessary that data does not change identity every time they change hands from one researcher to another. If this were not the case, it is hard to imagine how big data projects, or any project that requires a collaborative use of data, could be possible. I have in mind here projects where those who collect the data may be very far removed from, and have very different interests from, those many scientists who make use of their data. Leonelli discusses examples of this, with a different analysis, in (2016, chap. 1).

The analysis of data defended here differs in a second way from a relational account: re-identification occurs only when we change how we classify what we have recorded, but not every time data is put to a new use; consequently, data do not change often. Some circumstances that precipitate re-identification may include substantive theoretical change—such as a shift toward a heliocentric universe—or some other change in our understanding of the world or our instruments. A paradigm shift is an example of the kind of change required for re-identification, though milder shifts in understanding may also be sufficient, provided they result in new taxonomic distinctions. Perspectives allow us to talk about these kinds of data changes without invoking the kind of radical upheaval associated with paradigm shifts.

4 A third view of data

This case study suggests we need a different way of thinking about data, specifically that we should define them in terms of records.[3] I hinted at this at several points and also made appeal to something called a perspective. I believe to make sense of historical cases of data, an account needs to clarify both data-as-records and perspectives and the rest of this paper will take some first steps toward providing this clarification.

4.1 Data as records

The records associated with data can be of events, objects, behaviours, processes, plant specimens (Strasser 2012), any number of things that have been observed. For the case study discussed above, the records are records of astronomical events and objects. Scientists, we might say generally, use data to record, share, and carry specific details about a part of the world. In being used this way, data provide a means for scientists to work with a great deal of empirical content that would be impossible to aggregate otherwise. The function of data to record empirical events or other objects

[3] The analysis defended here is entirely concerned with the scientific data associated with scientific observation and experiment. There may very well be other forms of evidence, such as legal evidence used in legal systems, and other types of records and observations; The remarks in this article are not intended to address observation, records, and evidence beyond the scientific context. There may of course be important similarities, but there will likely be important differences as well.

of observation is of primary importance because, without this, they lack even the potential to function as evidence.

Using data as records, I contend, invariably requires human judgement. And if there is judgement involved, there is reason for thinking that data depend upon the epistemic context in which they feature. This appeal to judgement is necessary because it is otherwise impossible to establish which observations qualify for a data set; in other words, we must specify how to choose what to record, what it is we are recording, and how to group the records. The act of specifying, or classifying, what it is we are recording is particularly crucial here.

Specifying what observations to record, or what information to collect, is not trivial. Any given experiment can present to a researcher any number of things to record and only through judgement is it possible to extract data from this multitude. For instance, it is usually not the case that scientists would record whether they wore a watch, the size of the lab bench, what was for lunch, or the colour of their lab coats. It is impossible to record every such detail and scientists must make deliberate, informed choices about what is and is not important information. Those choices are informed by knowledge and understanding of, for example, theory, experimental techniques, or instruments. To return to a previous illustration, when melting lead, one might record the temperature because atomic theory predicts at what temperature lead melts. It is therefore an interesting study to compare what theory predicts with the world, which is done best, in this case, by comparing the actual temperature at which lead melts with what theory predicted. To do this, I need to know how to identify lead, how to melt it, how to measure its temperature as it melts, and how to read and record the reading on a thermometer.

In making a choice about what to record, scientists are classifying what it is they are recording; that is, what type of event, observation, or process their data represent. When I melt lead and record its temperature, I have provided a deceptively complex classification. Just in claiming something is lead, I am specifying that the substance is a type of metal, that it has certain logical relationships with other metals, different relationships with non-metal elements, reacts in certain ways in the presence of water or electricity, takes certain forms in progressive stages (its melting point is lower than its boiling point), and more. Equally complex things can be said of temperature. I do not have to have all this in mind, of course, and a lab technician might need to know very little about lead in order to melt it (perhaps just which drawer it's stored inside). However, if I do indeed have a set of data about melting lead, then there are a number of things that it is possible to say about what that data set represents, examples of which I have just listed. It is, therefore, very important that we have correctly classified what it is a particular datum represents.

A consequence of classifying what is recorded is being able to determine when two events are the same. If this were not the case, it would be unclear which data would belong in which set, and therefore what connections there are between data, how data can be used evidentially, and especially *what data we have*. There are criteria for justifying the treatment of two events as the same. Criteria, in this sense, are related to justification. To return to the lead example, I can justify using two measurements when my samples are of similarly pure lead that I heat using the same burner and measure using an accurate thermometer. Justification is important because scientists

must establish that their data sets of records of the same thing. This is particularly salient a problem in the case of unusual readings. Unusual readings suggest something has gone wrong. Say I have an outlier, a number that is much farther from the average temperature reading than my other recordings. If someone challenges me on this number, I need to be able to justify its inclusion and I do so by appealing to criteria that establish the outlier is the same type of event as the other recordings. I need to be able to say things like "I used the same thermometer and the same lead sample under the same conditions." Without this kind of justification, there is reason to doubt that my outlier can be included in my data set. And with this doubt comes uncertainty about what the data are records of because to make a record, one must know what it is one is recording.

Considering data collection as records of measurement or information-gathering makes more explicit the role of judgement in collection. What we measure or what information we gather is a selective choice; we do not record and measure all possible parameters or information, just a subset. Furthermore and more importantly for my discussion, in judging what to record, those who collect data distinguish the phenomenon of interest from others that are irrelevant for their purposes. Millikan had to be able to distinguish falling, charged oil drops from all other phenomena surrounding his experiment, including oil drops that were not charged. I am considering this selective recording a judgement because Millikan could have chosen to record other things (though that information would have been useless for his study) and because he required extensive knowledge about electrons, oil drops, and his instruments in order to make the kinds of records he did.

This view of data is in keeping with some philosophy of experiment. Hacking's (Hacking in Pickering 1992, chap. 2) taxonomy of the elements of laboratory experimentation shows the tight interconnections between different types of theory, research, experimental equipment, and the products of an experiment. When one element changes, typically another element will require adjustment. We can infer from his account that any changes in the data will require adjustment in other parts of laboratory experiment, be it theory, hypothesis, or explanation of the experimental equipment. And conversely, any changes in the laboratory may invoke changes in data. I am not restricting my discussion of data to the laboratory, as Hacking does, but we can say more generally that the stability of data requires a stability in other parts of the empirical inquiry, i.e. the epistemic context.

These records, as the case study illustrated, are subject to certain kinds of changes as the epistemic context changes. Because of these changes, it is fruitful to think of data as perspectival.

4.2 Perspectival data

We might say that data, because of their dependence upon the epistemic context, are perspectival. Perspectivism, as Giere (2006) articulates it, is about representations, especially models. But the term perspectivism is just as apt for data because data are representational and because perspectivism also emphasizes the contextual nature of scientific practice, which is a feature that data have. Another reason to invoke

perspectivism concerns how scientists make judgements. Recent work by Massimi (2012) suggests that we conceive of perspectivism as the epistemic context in which scientists work. Scientific work is always situated within a network of beliefs that should be reliably justified and coherent. She writes "Justified-belief-attribution is always perspectival and contextual: it has to do with the way each belief fits into the agent's epistemic perspective" (2012, p. 48).

Data are part of the epistemic context because scientists make judgments based upon their beliefs. And because collecting data is associated with a judgment, data collection should be cast as a perspectival activity that partly forms a perspective. If this is right, then these judgments are not merely informed by the epistemic contexts in which scientists work: they in fact partly constitute this epistemic context.

Another reason to invoke perspectivism here is to distinguish between context-dependence and theory-dependence.[4] Theory-dependence suggests a high level of dependence on a particular theory, i.e. on a very specific and abstract body of knowledge, such as atomic theory. Hacking (Hacking in Pickering 1992, p. 45) calls this systematic theory, as opposed to topical hypotheses or models of the laboratory apparatus. It seems to me that we often speak of objects or particles without committing to an abstract or systematic theory or while remaining agnostic about which theory we endorse. Consider again Millikan and Franklin. The atomic theories each endorsed are not exactly the same, yet there is a sense in which they were interested in the same particles. If electrons were theory-dependent, it seems to me that Millikan and Franklin could not be interested in the same particles.

Data, although they depend upon the epistemic context, do not depend upon theory heavily. If they did, data could neither arbitrate between theoretical claims nor support or refute such claims. The practice of collecting data sets, analyzing them, and using them evidentially would then be very mysterious or even pointless. So data must have some level of independence from theory, which is part of what Bogen and Woodward (1988) too such pains to show. Hacking (Hack also argues, but in the context of observation more generally, that there must be enough independence from theory to provide empirical constraints.

The kind of epistemic context upon which data do depend—and of which data are a part—is more fundamental than theory. This context includes a range of knowledge, such as knowledge of instruments, experimental technique, and especially knowledge classification. That is, how to classify what is observed. Scientists make judgments about classification in association with data-collection and use. Such a judgment might be "this is a meteorological object," or "this is a lead sample." These judgments are perspectival, I suggest, because they are intimately connected to the understanding scientists have and this understanding is distinct from theory. One can, for instance, judge an object in the sky to be a comet and they might make this judgment independently of modern astronomical theory. And yet if the epistemic context changes sufficiently, scientists would make, or might make, different judgments. Instead of judging comets as meteorological, we now judge them to be astronomical.

[4] For some discussion of the issue of theory-ladenness, see for example (Kordig 1971; Brewer and Lambert 2001; Schindler 2011).

4.3 Some previous problems avoided

Recall that accounts of data must confront identity and stability problems. That is, the account must explain what data are such that they can be evidential and they must be able to specify when and why a set of data is the same. Treating data as records avoids both. The first problem is avoided because this definition does justice to the potential that data have to be evidentially useful while not overplaying their use as evidence. It is clear why a scientist would collect data: because they are records of worldly events against which theories or models can be tested. And if this is what data are, then it is no longer mysterious why scientists would find them worth collecting.

The issue of stability is avoided because data remain stable to the extent that records are stable. However, this treatment may bring to mind the representational view that Leonelli criticises (2016, pp. 73–74). The representational view is committed to the two things: (1) that data are mind-independent representations when reliably produced by the scientific method and (2) that the exclusive role of data is to support claims about phenomena.[5] Leonelli argues that data cannot be defined in this way. I am not attempting to defend a representational view with these two commitments. I have suggested data are not just evidence, which conflicts with (2). I have also tied data to the epistemic context, which conflicts with one.

5 Conclusion

In this paper, I sought to address what data are and what relationship they have to the epistemic context in which scientists work. A case study from the history of astronomy motivated the view I proposed, which has the advantage of accommodating the empirical contribution data make while also giving due consideration to the role of human knowledge and understanding. To do this, I argued that, contra relational accounts, data are generally highly stable across different inquiries and across material forms. Data have this stability because they are records. I then argued (1) that data are not stable in the face of certain kinds of scientific change and (2) that we can think of data as perspectival in the sense that they are part of the epistemic context. This contextual dependence comes from the distinctions scientists make about what they record. One upshot of thinking about data in this way is that it shows how even old data can be repurposed and improved to reflect the contemporary epistemic context and help scientists investigate contemporary research questions. This proposal also gives credence to the profound changes in science and the surmountable difficulties associated with navigating the ancient sciences and our changing understanding of the world.

Acknowledgements This article comes out of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement European Consolidator Grant H2020-ERC-2014-CoG 647272 *Perspectival Realism. Science, Knowledge, and Truth from a Human Vantage Point*). I would like to thank Michela Massimi and

[5] Woodward (2011, p. 166; 1989, pp. 393–94) and Bogen and Woodward (1988, p. 305) frequently make this claim.

Alasdair Richmond for reading earlier versions of this article. Audiences at the 2017 Annual UK History and Philosophy of Science Workshop provided helpful feedback.

Compliance with ethical standards

Conflict of interest There are no interests, apart from those mentioned in the Acknowledgements.

References

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review, 97*(3), 303–352.

Bogen, J., & Woodword, J. (2003). Evading the IRS. *Poznań Studies in the Philosophy of the Sciences and the Humanities, 20,* 223–256.

Brewer, W. F., & Lambert, B. L. (2001). The theory-ladenness of observation and the theory-ladenness of the rest of the scientific process. *Philosophy of Science, 68*(S3), S176–S186.

Clark, D. H., & Richard Stephenson, F. (1977). *The historical supernovae* (p. 233). Oxford: Pergamon International Library of Science, Technology, Engineering and Social Studies. £8.50. *Bulletin of the School of Oriental and African Studies*, *41*(3), 627–628.

Franklin, A. D. (1981). Millikan's published and unpublished data on oil drops. *Historical Studies in the Physical Sciences, 11*(2), 185–201.

Giere, R. (2006). *Scientific perspectivism*. University of Chicago Press.

Goodstein, J. R., Gunns, A. F., & Underleak, A. (1977). *The robert andrews millikan collection at the California Institute of Technology: Guide to a microfilm edition*. Pasadena: California Institute of Technology.

Green, D., & Stephenson, F. (2008). "The historical supernovae". In K. Weiler (Ed.), *Supernovae and gamma-ray bursters*. Springer.

Kordig, C. R. (1971). The theory-ladenness of observation. *The Justification of Scientific Change* (pp. 1–33). Dordrecht: Springer.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Cambridge: Harvard University Press.

Leonelli, S. (2009). On the locality of data and claims about phenomena. *Philosophy of Science, 76*(5), 737–749.

Leonelli, S. (2012). Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Studies in the History and the Philosophy of the Biological and Biomedical Sciences: Part C, 43*(1), 1–3. https://doi.org/10.1016/j.shpsc.2011.10.001.

Leonelli, S. (2013). Integrating data to acquire new knowledge: Three modes of integration in plant science. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 44*(4), 503–514. https://doi.org/10.1016/j.shpsc.2013.03.020.

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science, 82*(5), 1–11.

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. Chicago: University of Chicago Press.

Massimi, M. (2012). Scientific perspectivism and its foes. *Philosophica, 84,* 25–52.

Pickering, A. (1992). *Science as practice and culture*. Chicago: University of Chicago Press.

Rheinberger, H. (2011). Infra-experimentality: From traces to data, from data to patterning facts. *History of Science, 49*(3), 337–348.

Schindler, S. (2011). Bogen and Woodward's data-phenomena distinction, forms of theory-ladenness, and the reliability of data. *Synthese, 182*(1), 39–55.
Strasser, B. J. (2012). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(1), 85–87.
Thorsett, S. E. (1992). Identification of the Pulsar PSR1509—58 with the 'guest Star' of AD 185. *Nature, 355,* 717–719.
Wang, Z., Qu, Q., & Chen, Y. (1997). Is RX J1713. 7-3946 the remnant of the AD393 guest star? *Astronomy and Astrophysics*, *318*, L59–L61.
Woodward, J. (1989). Data and phenomena. *Synthese*, *79*(3), 393–472.
Woodward, J. (2010). Data, phenomena, signal, and noise. *Philosophy of Science, 77,* 792–803.
Woodward, J. (2011). Data and phenomena: A restatement and defense. *Synthese*, *182*(1), 165–179.